DOCUMENT RESUME

ED 447 194 TM 032 083

AUTHOR McMillan, James H.

TITLE Examining Categories of Rival Hypotheses for Educational Research.

PUB DATE 2000-04-25

NOTE 9p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).

PUB TYPE Opinion Papers (120) -- Reports - Descriptive (141) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Classification; *Educational Research; *Hypothesis Testing; *Test Interpretation; *Validity

ABSTRACT

Implicit in the work of S. Huck and H. Sandler (1979) is the idea that the concept "rival hypotheses" refers to some kind of alternative explanation. Rather than being threats to internal validity, rival hypotheses, in their view, are interpretations that differ from those of the researcher. This paper broadens the idea to include any interpretations or explanations that are important in understanding and using results, whether these are inconsistent with those of the researcher or simply noted as plausible. Huck and Sandler present 20 categories of rival hypotheses, going beyond the classic list of 7 threats to internal validity (Campbell and Stanley, 1963). This paper organizes categories of threats, dividing them into threats to external validity (or generalizability) and threats to internal validity. Three categories within internal credibility correspond to statistical conclusion, relationship conclusion, and causal conclusion explanations. These three categories are further subdivided, as shown in a table of potential threats to validity for experimental and non-experimental designs. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 447 194

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
J. McMillan
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

Examining Categories of Rival Hypotheses for Educational Research

James H. McMillan
Virginia Commonwealth University

TM032083

Paper presented at the American Educational Research Association Annual Meeting, New Orleans, April 25, 2000.

BEST COPY AVAILABLE

It is indeed a sobering exercise to consider modifications or changes to the seminal work of Don Campbell, Julian Stanley, and Thomas Cook, as well as to an excellent application of this work by Huck and Sandler (1979). But after teaching principles of experimental validity for over twenty years, and considering some recent recommendations of professional organizations concerning research design and the use of statistics, I do think that a different organization of many of the ideas may help students understand why application of these principles is crucial to conducting credible research and being an informed consumer of research.

To provide a context for my recommendations I would point out that while Cook and Campbell (1979) acknowledged that many so-called "threats" to experimental validity could be applicable to nonexperimental, and well as experimental research, such distinctions in current conceptualizations do not emphasize this difference very much. With some exceptions, internal and external validity are typically presented in the context of experimental design, and sadly in my view, many authors still present only the original categories of threats to internal and external validity explicated by Campbell and Stanley in 1963. But when you consider the addition of threats to statistical conclusion and construct validity, it is clear that both experimental and nonexperimental propositions can be evaluated. External validity, on the other hand, has been applicable to both types of research.

Also, I agree very much with David Krathwohl (1998) that the use of the terms "internal validity" and "external validity," while justified from what is meant by the dictionary definition of validity, was unfortunate because of confusion with test validity. Cook and Campbell did not help matters any when they used the term "construct validity," because these two words are also widely used in measurement. In addition, some of the original labels given for threats to validity are either imprecise or misleading (e.g., testing, history and selection). For these reasons, terminology is a major issue with students when teaching these concepts.

Implicit in the work of Huck and Sandler (1979) is the idea that the concept "rival hypothesis" refers to some kind of alternative explanation. In their view, a rival hypothesis is not the same as threat to internal validity. Rather, rival hypotheses are interpretations that differ from those of the researcher. I would broaden this idea to include any interpretations or explanations that are important in understanding and using results, whether these are inconsistent with those of the researcher(s) or simply noted as plausible. That is, the idea of rival should not be limited to something that "rivals" stated interpretations and conclusions of the researcher, but

should be inclusive of whatever explanations are most reasonable or plausible, given limitations as suggested by internal validity threats.

Huck and Sandler (1979) present 20 categories of rival hypotheses. They admonish readers to go beyond the classic list of seven threats to internal validity of Campbell and Stanley, and I agree with them that this list of seven alone is insufficient in identifying important alternative explanations. They defer threats to external validity to other authors. While I have used the Huck and Sandler approach often and value very much the scenarios they present and analyze, I find that there are some additional "threats" that need to be considered, especially given the work of Cook and Campbell and more recent thinking about statistical significance.

I will begin my definition of what is meant by such terms as "internal validity" and "rival hypotheses," then organize possible threats into categories that make sense to me when considering applications to both experimental and nonexperimental research. To keep matters relatively strait forward, I am concerned here with quantitative studies. Though the notion of credibility of explanations and interpretations clearly concern qualitative as well as quantitative research, and some so-called threats to validity are applicable to qualitative studies (e.g., researcher bias, observer error), I will leave it to others to try to integrate the language of these two traditions. My experience is that increased understanding results when these two major types of research, and applicable principles related to credibility, and presented separately.

To begin, it makes most sense to me to think about internal validity as the reasonableness and credibility of findings, claims, propositions, explanations, interpretations, and conclusions *made within the local context of particular study*. This definition is not restricted to causal interpretations or conclusions. Internal validity could even be renamed *internal credibility* to lessen confusion with test validity. When we think about threats to internal credibility we are essentially focused on rival hypotheses, though even the term hypotheses in this context conjures up images of experiments. What we are looking for in the threats are rival explanations to the propositions made on the basis of the data. We are also interested in the *plausibility of the threats*. Here I think it is important to stress to students that while a threat may be **possible**, that doesn't mean it is **plausible**. To put it differently, just because a given threat is not controlled by the design doesn't mean that the threat is plausible or that the design is necessarily weak. Huck and Sandler (1979) make the same argument. Furthermore, students need to understand that to be "plausible" in experiments, threats must create a condition where one group is influenced

more than another one (e.g. if both groups in an experiment take a pretest testing is probably not a threat because its affect is very similar if not the same for both groups).

With these thoughts in mind, I believe it makes sense to think about five types of propositions, then organize possible threats among these categories for both experimental and nonexperimental designs. I have taken the liberty of renaming some familiar threats, to adding some new ones, to using some suggested by Huck and Sandler, and to make a fundamental change by including threats to construct validity under internal credibility. This last change is made because of the importance of confounding and the enhanced clarity of conceptualizing threats to external validity as only those that involve generalization of findings to other situations and individuals (i.e., generalizing *to*). I find that this more narrow definition of external validity (which could be renamed simply *generalizability*) is conceptually clear and straightforward. If threats to Cook and Campbell's construct validity are thought of as a form of generalizing, then there is a great deal of confusion. Also, the idea of confounding is essentially a concern for interpreting the findings of a particular study, not something that makes intuitive sense from a generalizability standpoint. Confounding is very important to the reasonableness of a particular inference, explanation, or conclusion. Huck and Sandler seem to agree by including "treatment confound" as one of their proposed categories of rival hypotheses.

My organization is illustrated in Table 1. I have first divided the threats into two categories – those pertaining to external validity as one, and the rest of them (internal credibility), as a second. There are three categories within internal credibility to correspond to major types of conclusions and explanations that are made: statistical, relationship, and causal. Statistical conclusion includes threats posed by Cook and Campbell, instability and statistical threats of Huck and Sandler, and the addition of effect size. Relationship conclusion threats are mostly related to nonexperimental designs using either correlations or comparisons to study possible relationships, while causal conclusion threats mostly focus on experimental design.

While most of the individual threats in Table 1 are familiar, two deserve some further explanation. It is clear that understanding the difference between statistical significance and magnitude of relationships and differences is essential to accurate interpretations and conclusions. I have selected the term "effect size" as a category to bring attention to this principle, primarily because it is a term that is being used more and more frequently. In fact many journals now require effect size information. A second threat that I have included is called

"treatment replications." This threat refers to the number of times treatments are replicated, and whether the treatments are replicated independently for each subject. Similar to what Huck and Sandler call "treatment confounds," and what Cook and Campbell term construct validity, a common problem in applied educational experiments is whether or not the treatment is replicated independently for each "subject." From a statistical standpoint this relates to unit of analysis, but from a research design standpoint, single replications in group settings invite many alternative explanations because it is easy and probable that some kind of group interaction effect, or event, or peculiarity in the treatment, influences the outcome (e.g., the "lawnmower effect") (see McMillan, 1999, for more detail on treatment replications as a possible threat).

From a practical perspective, the list of threats in Table 1 is too long. What is needed is further thought about how such a list could be synthesized into something more reasonable, while at the same time providing the level of specificity needed to cover important aspects of research that we know from experience need to be considered in making reasonable and accurate interpretations and explanations. I'll look forward to seeing and using such a synthesis as well as others' ideas about different ways of categorizing threats or rival hypotheses.

References

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago: Rand-McNally.

Huck, S. W., & Sandler, H. M. (1979). Rival hypotheses: Alternative interpretations of data based conclusions. New York: Harper & Row.

Krathwohl, D. R. (1998). *Methods of educational and social science research: An integrated approach* (2ed). New York: Longman.

McMillan, J. H. (1999). Unit of analysis in field experiments: Some design considerations for educational researchers. ERIC no. ED428135.

Table 1

Categorizing Potential Threats to Validity for Experimental and Nonexperimental Designs[1]

Internal Credibility

Statistical Conclusion	Experimental	Nonexperimental
▪ Type I errors	✓	✓
▪ Type II errors	✓	✓
▪ Violated assumptions of statistical tests	✓	✓
▪ Fishing and the error rate problem	✓	✓
▪ Reliability of measures	✓	✓
▪ Reliability of treatment implementation	✓	
▪ Random irrelevancies in the setting	✓	✓
▪ Random heterogeneity of respondents	✓	✓
▪ Effect size	✓	✓
▪ Distorted graphics	✓	✓
Relationship Conclusion		
▪ Restricted range	✓	✓
▪ Outliers	✓	✓
▪ Curvilinear data		✓
▪ Homogeneity of respondents	✓	✓
▪ Correlation and causality	✓	✓
▪ Ambiguity about the direction of cause		✓
▪ Multicollinearity		✓
▪ Unaccounted-for variables		✓

[1] A check mark indicates a possible threat.

Causal Conclusion

• Extraneous events (history)	✓	
• Internal events (history)	✓	
• Maturation	✓	✓
• Pretesting (testing)	✓	
• Instrumentation	✓	✓
• Statistical regression	✓	
• Lack of random assignment (selection)	✓	
• Matching (selection)	✓	
• Subject attrition (mortality)	✓	
• Order effect	✓	✓
• Interactions with selection	✓	
• Diffusion of treatments	✓	
• Compensatory rivalry	✓	
• Resentful demoralization	✓	
• Experimenter bias	✓	✓
• Observer, recorder or rater bias	✓	✓
• Hawthorne effect	✓	✓
• Treatment replications	✓	
• Demand characteristics	✓	✓
• Inadequate preoperational explication	✓	✓
• Mono-operation bias	✓	✓
• Mono-method bias	✓	✓
• Hypothesis guessing	✓	✓
• Evaluation apprehension	✓	✓
• Experimenter expectations	✓	✓
• Confounding constructs and levels of constructs	✓	✓
• Interaction of different treatments	✓	
• Interaction of testing and treatment	✓	
• Novelty	✓	

Generalizability

Population

• Generalizing subject characteristics to others	✓	✓
• Generalizing within	✓	✓

Ecological

• Nature of measures	✓	✓
• Interaction of selection and treatment	✓	
• Interaction of setting and treatment	✓	
• Interaction of history and treatment	✓	
• Inadequate sampling	✓	✓

END

U.S. Dept. of Education

Office of Educational
Research and Improvement (OERI)

ERIC

Date Filmed
April 26, 2001